

Word Count: 4477
Tables:2
Figures: 2

Mental Health Computerized Adaptive Testing

Robert D. Gibbons¹
David Weiss²
David Kupfer³
Ellen Frank³
Victoria Grochocinski³
Dulal Bhaumik¹
Angela Stover³
R. Darrell Bock¹

April, 2005

¹Center for Health Statistics, University of Illinois at Chicago

²Department of Psychology, University of Minnesota

³Western Psychiatric Institute, University of Pittsburgh

Acknowledgements: This work was supported by NIMH grant R01-MH66302.

Corresponding Author:

Robert D. Gibbons Ph.D.
Director, Center for Health Statistics
University of Illinois at Chicago
1601 W. Taylor
Chicago IL 60612
(312) 413-7755 (phone)
(312) 996-2113 (fax)
e-mail:rdgib@uic.edu

ABSTRACT

Context: Mental health measurement relies upon antiquated approaches to the analysis, design and administration of psychiatric instruments. **Objective:** This study examines the utility of the combination of Item Response Theory (IRT) and computerized adaptive testing (CAT) for the calibration, administration, and scoring of psychiatric measurement systems. **Design:** To illustrate the application of IRT-based CAT to a problem in psychiatric measurement, we selected the Mood and Anxiety Spectrum Scales (MASS). The MASS consists of 626 items covering mood spectrum, panic-agoraphobic spectrum, obsessive-compulsive spectrum, and social phobic spectrum. Based on a balanced incomplete block design, 36 forms each consisting of 154 items were administered to 800 participants who were in outpatient treatment for a mood or anxiety disorder at Western Psychiatric Institute and Clinic (WPIC, Pittsburgh, PA). **Results:** A bi-factor IRT model provided an excellent fit to the observed data, and 90% of the items were included in the item bank. Simulated CAT administration for subjects that had completed the entire test, revealed a 96% average reduction in the number of items administered for CAT (24 items) relative to traditional scale administration (615 items). The correlation between total MASS score and CAT-MASS score was $r=93$, indicating excellent agreement between the full test and reduced test administrations of the MASS. **Conclusions:** IRT-based CAT provides a new paradigm for mental health measurement that goes well beyond traditional approaches to psychiatric measurement. Instead of designing scales of a fixed length and small number of items by which to evaluate subjects of various levels of impairment, potentially unlimited numbers of items can be used to construct an item bank, and a small set of the most relevant items for a given subject can be administered with large gains in test information as well as a dramatic reduction in administration time.

INTRODUCTION

Mental health research relies heavily on antiquated systems of measurement. The construction of traditional mental health scales is based largely on subjective judgment, and at best, application of methods from classical test theory to determine a scale's psychometric properties. In almost all cases, the level of impairment of the subject is determined by a simple total of the individual symptom scores. Remarkably, there have been enormous developments in educational measurement along these lines (*i.e.*, *Item Response Theory - IRT*), however, little if any of these new developments have found their way into the mainstream of mental health measurement. Furthermore, in all cases, each subject must be measured on exactly the same items (*e.g.*, symptom scores) in order to be evaluated on the construct that the scale was designed to measure (*e.g.*, depression).

Ideas of adaptive testing in which different subjects may receive different items that are targeted to their level of impairment have largely not been considered in mental health research. As an analogy, consider a mathematics test covering arithmetic through calculus. Based on an individual's responses to a small screening set of items, we can determine a provisional estimate of the individual's ability and target further items to that ability level. Indeed, we can learn very little about an individual who is competent at calculus by administering basic addition items. Conversely, we can learn very little about a mathematically remedial individual by administering calculus items. By employing a statistically-based model of psychometric measurement (*e.g.*, an IRT model), we can calibrate both items and individuals, and more efficiently identify subsets of suitable items for a particular individual. This general idea is referred to as computerized adaptive testing (CAT) and has immediate application possibilities in mental health research. For example, a large inventory of depression items can be administered adaptively, such that initial estimates of the subject's level of impairment can be used to select the most appropriate additional items to administer for that subject.

The net result is that large "banks" of items can be established, calibrated on a common scale or scales, and administered adaptively such that any individual subject is only rated on a much smaller number of symptoms without loss of measurement precision. Through the use of an IRT

model, statistically comparable estimates of impairment can be obtained for different subjects who were rated on potentially different items (*e.g.*, depressive symptoms).

A complication of applying IRT to mental health measurement problems is that unlike traditional ability testing (*e.g.*, mathematics achievement) which are inherently unidimensional, mental health measurement scales are inherently multidimensional. Although multidimensional IRT models are available (Bock and Aitkin, 1981; Bock, Gibbons and Muraki, 1988) they have not been well studied in the context of adaptive testing. Note that one of the primary reasons for the multidimensionality of mental health measurement scales is that the items are often sampled from multiple domains (*e.g.*, various mood disorders), and it is common to observe excess correlation within domains than would be expected under the conditional independence assumption of a unidimensional IRT model. To this end, Gibbons and Hedeker (1992) developed an “item bi-factor” model, which allows each item to load on a primary dimension (*e.g.*, depression) and one subdomain (*e.g.*, sleep disturbance). In the context of mental health measurement, the advantage of the bi-factor model is that it yields a measure of overall impairment that can be the focus of adaptive testing.

In the following sections, several foundational components that lead up to IRT-based CAT of binary and ordinal mental health measurement instruments are described and illustrated using an example dataset.

The Logic of IRT

For those already familiar with traditional methods of educational and psychological testing, an understanding that classical and IRT methods of scoring tests are based on entirely different premises is crucial. The difference is clarified by the following analogy. Imagine a track and field meet in which ten athletes participate in men’s 110-meter hurdles race and also in the men’s high jump. Suppose that the hurdles race is not quite conventional in that the hurdles are not all the same height and the score is determined, not only by the runner’s time, but also by the number of hurdles successfully cleared, *i.e.*, not tipped over. On the other hand the high jump is conducted in the conventional way: the cross bar is raised by, say, 2 cm increments on the

uprights, and the athletes try to jump over the bar without dislodging it.

The first of these two events is like a traditionally scored objective test: runners attempt to clear hurdles of varying heights which is analogous to questions of varying difficulty that individuals try to answer correctly in the time allowed. In either case, a specific counting operation measures ability to clear the hurdles or answer the questions.

On the high jump, ability is measured by a scale in millimeters and centimeters on the upright and the highest scale position of the cross bar the athlete can clear. IRT measurement uses the same logic as the high jump. Test items are arranged on a continuum at certain fixed points of increasing difficulty. The individual attempts to answer items until he can no longer do so correctly. Ability is measured by the location on the continuum of the last item answered correctly. In IRT, ability is measured by a scale point, not a numerical count.

These two methods of scoring the hurdles and the high jump, or their analogues in traditional and IRT scoring of objective tests, contrast sharply: if hurdles are arbitrarily added or removed, number of hurdles cleared cannot be compared with races run with different hurdles or different numbers of hurdles. Even if the percent of hurdles cleared were reported, the varying difficulty of clearing hurdles of different heights would render these figures non-comparable. The same is true of traditional number-right scores of objective tests: scores lose their comparability if item composition is changed.

The same is not true, however, of the high jump or of IRT scoring. If the bar in the high jump were placed between the 2 cm positions, or if one of those positions were omitted, height cleared is unchanged and only the precision of the measurement at that point on the scale is affected. Indeed, in the standard rules for the high jump, the participants have the option of omitting lower heights they feel they can clear. Similarly, in IRT scoring of tests, a certain number of items can be arbitrarily added, deleted or replaced without losing comparability of scores on the scale. Only the precision of measurement at some points on the scale is affected.

This property of scaled measurement, as opposed to counts of events, is the most salient

advantage of IRT over classical methods of educational and psychological measurement. From this analogy, it should be clear that the current “state of the art” in mental health measurement is completely based on traditional or classical methods and not IRT.

METHODS

Computerized Adaptive Testing

1. Basic Concepts

In an adaptive test, items are selected during the process of test administration for each individual being tested. Adaptive tests are designed to allow the test administrator to control the precision of a given measurement and to maximize the efficiency of the testing process.

The characteristics of an adaptive test include:

1. *A precalibrated bank of test items.* To create an adaptive test, items must previously be administered to a group of individuals, and item difficulty (symptom severity), and discrimination (ability to discriminate high and low levels of impairment) must be computed for each symptom item.
2. *A procedure for item selection.* Because items are selected based on an individual’s previous responses, items must be scored as they are administered. The next item (or item subset) to be administered is then based on how the individual answered all previously administered items.
3. *A method of scoring the test.* In an adaptive test, not only must the items be scored as they are administered, but the subject must be scored as well after each item is administered.
4. *A procedure for terminating the test.* In contrast to a conventional test, the number of test items is not fixed in an adaptive test. The test is terminated when the uncertainty in the score is small.

Research since the 1970s has shown that these four characteristics of adaptive testing procedures are most easily achieved using IRT procedures (*e.g.*, Kingsbury & Weiss, 1980, 1983; McBride & Martin, 1983). Within the last two decades, commercially available software for implementing IRT-based CAT has been developed and upgraded (Assessment Systems Corporation, 1987, 2001), and tests such as the Graduate Record Examination (GRE) have become CATs. Research shows that adaptive tests are more efficient than conventional tests (*e.g.*, Brown & Weiss, 1977; McBride & Martin, 1983). That is, in an adaptive test a given level of measurement precision can be reached much more quickly than in a test in which all individuals are administered the same items. Typical adaptive tests result in a 50% average reduction in number of items administered, and some reductions in the range of 80% to 90% have been reported, with no decrease in measurement quality (Brown & Weiss, 1977

IRT measurement models also includes numerous models that are applicable to personality instruments that are not dichotomously scored (*e.g.*, Andrich, 1978a, 1978b, 1988; Muraki, 1990; Samejima, 1969). Research has demonstrated that the IRT family of models can be meaningfully applied to the measurement of attitudes and personality variables (*e.g.*, Reise & Waller, 1991), including CAT (*e.g.*, Baek, 1997, Dodd, De Ayala, & Koch, 1995).

IRT Models for Psychiatric Measurements

Most typical applications of IRT have assumed unidimensional item sets, *i.e.*, items for which the responses could be accounted for by a single attribute for each subject. However Bock and Aitkin (1981) and Bock, Gibbons and Muraki (1988) extended the IRT model to the multidimensional case, where each item is related to one or more underlying latent dimensions, traits, or constructs of interest. In general, however, multidimensional CAT is problematic, because it involves selection of items that will increase precision of measurement simultaneously on all dimensions of interest. In part, however, this multidimensionality is produced by the sampling of items from multiple domains of an overall psychological construct. For example, in the measurement of life quality, items are selected from satisfaction domains such as satisfaction with family, income, neighborhood, etc. It is quite natural for such data to appear to be multidimensional, when in fact, they measure a unidimensional construct, *i.e.*, quality of life,

however, the items within domains are more highly correlated than items between domains.

A good alternative is based on a “bi-factor” model. The bi-factor solution constrains each item to have a non-zero loading on the primary dimension and a second loading on no more than one of domain factor (Holzinger and Swineford, 1937). It is plausible mental health measurement, where symptom items are often selected from measurement domains and can be related to the primary dimension of interest (*e.g.*, depression) and one subdomain (*e.g.*, anxiety).

Gibbons and Hedeker (1992) derived a bi-factor IRT model for binary response data, and Gibbons et.al. (2005) extended it for analysis of ordinal response data. Their estimation method permits any number of item domains, and provides a single estimate of the primary dimension the test was designed to measure that is suitable for CAT.

Psychiatric Measurement of Mood Disorders

Notwithstanding the enormous utility of the RDC and DSM III, III-R and IV in advancing our understanding of clinical course, biological bases and treatment of mood disorders, there is increasing recognition that these disorders rarely come in the pure and seemingly isolated prototypes described in the current nomenclature. Indeed, co-occurrence of depression, both with other Axis I disorders and with Axis II disorders, appears to be the rule. But this kind of comorbidity is only the tip of the iceberg. Observant clinicians are aware that patients describe myriad symptoms either not in the DSM criteria or not clustered as DSM indicates. To this end, Cassano and colleagues at the University of Pisa formed a U.S.-Italian collaborative group that has worked for five years to define the full “spectrum” of clinical features of mood and anxiety disorders and to create systematic assessment strategies for these features (Cassano et al., 1997; Frank et al., 1998). These spectrum assessments reflect a more inclusive perspective on psychopathology and, as such, may have great utility in refining our disorder concepts and our conceptualization of the relationships among them. A more inclusive perspective, in turn, offers the promise of treatment-relevant phenotypes or a priori identification of subgroups likely to exhibit a particular response to a particular treatment.

Preliminary data on the Spectrum Scales

Traditional psychometric properties of the Spectrum scales have been previously established (Cassano et al., 1999-a; Cyranowski et al., 2002; Dell’Osso et al., 2000; Fagiolini et al., 1999; Mauri et al., 2000; Sbrana et al., 2003; Sbrana et al., submitted). Test-retest and inter-rater reliability ranged from .89 to .99 and each showed concurrent validity when compared with established measures of comparable constructs. Excellent levels of agreement have been shown between the structured clinical interviews and the more efficient self-report measures: mood spectrum, $r=0.97$ (Dell’Osso et al., 2002-a); panic-agoraphobic spectrum, $r=0.94$ (Shear et al., 2001); obsessive-compulsive spectrum, $r=0.96$ and social phobia spectrum, $r=0.97$ (Dell’Osso et al., 2002-b). Since the instruments were co-developed in Italian and English, cross-cultural validity has also been demonstrated (Frank et al., submitted).

The mood and anxiety spectrum measures have been shown to have substantial clinical utility, displaying associations with unipolar depression treatment outcome (Frank et al., 2000), functional impairment (Shear et al., 2001), and bipolar disorder treatment outcomes (Frank et al., 2002; Rucci et al., 2003-a) that persist after controlling for traditional DSM diagnostic comorbidity (Cassano et al., 1997, 1999-b, 2004; Rucci et al., 2003-a).

RESULTS

Application: The Mood-Anxiety Spectrum Disorders Scale (MASS)

To illustrate the application of IRT-based CAT to a problem in psychiatric measurement, we selected the MASS. The MASS consists of 626 items and assesses mood spectrum, panic-agoraphobic spectrum, obsessive-compulsive spectrum, and social phobia spectrum (SCI-MOODS (161 items), SCI-PAS (114 items), SCI-OBS (183 items), SCI-SHY (168 items), respectively). A self-rating version of the MASS has been developed and validated to assess lifetime mood and anxiety spectrum conditions.

1. Balanced Incomplete Block Design

We derived an optimal balanced incomplete block design (BIB) that maximizes the number of pairings of 616 of the 626 individual items while minimizing the number of items administered to each subject (Cochran & Cox, 1957). The FastTEST Pro Testing System (Assessment Systems Corporation, 2000) was used to create 36 different forms of the test, each consisting of an optimal selection of 154 items extracted from the four MASS instruments. Data from 800 participants were collected.

2. Demographic Characteristics

Of the total 800 subjects, 567(71%) were women. The mean age was 38.3 years (SD = 11.6, range = 18 - 66), and the mean education level was 14.3 years (SD = 2.7), which are both typical of outpatient populations in mental health. Table 1 displays the demographic composition of our sample. The proportion of African Americans in the MHCAT study is twice the proportion in the greater Pittsburgh area (26% in the study vs. 13% in the Pittsburgh area). The study participants had a broad range of computer experience, and it is notable that 11% had never used a computer prior to the study.

Over 95% of participants (n = 765) endorsed a preference for answering mental health questions via computer rather than paper-and-pencil. The FastTEST Pro interface was rated as good or excellent by 90% (n = 722) of MHCAT participants and 85% (n = 680) reported that the program was easy or very easy to complete. A majority of participants (85%, n = 679) also indicated that they felt comfortable or very comfortable answering personal questions via computer. Participants were also asked how well the questions described their experiences with mental health problems and 70% (n = 560) felt their experiences were described very much or a great deal. In fact, 62% (n = 492) felt that a lot or a great deal of insight into individual mental health symptoms could be gained by their clinician or physician if the MASS items were reviewed.

Table 1. Demographic Characteristics.

Race	Total (%)	Computer Usage	Total (%)
Caucasian	545 (68%)	Never	91 (11%)
African American	209 (26%)	Less than 1x/month	86 (11%)
More than 1 Race	25 (3%)	Monthly	56 (7%)
Hispanic	12 (1%)	Weekly	95 (12%)
Refused	9 (1%)	Daily	470 (59%)
American Indian	6 (LT 1%)		
Asian	6 (LT 1%)	Format Preference	
		Paper-and-pencil	32 (4%)
		Computer	765 (96%)

All participants were in outpatient treatment for a mood or anxiety disorder at Western Psychiatric Institute and Clinic (WPIC, Pittsburgh, PA). Medical charts from all patients were reviewed for diagnostic information. Table 2 shows the frequencies and percentages of diagnostic categories, which are not mutually exclusive (e.g., depression can co-occur with an anxiety disorder, etc.).

Table 2. Diagnostic Characteristics.

Diagnosis	Total (%)	Diagnosis	Total (%)
Depression (Single, Recurrent, NOS)	569 (71%)	Past Alcohol Dependence	66 (8%)
Anxiety (GAD, Panic, Social, NOS)	308 (39%)	Obsessive-Compulsive Disorder	56 (7%)
Bipolar (I, II, NOS)	176 (22%)	Eating Disorder (Bulimia, Anorexia, NOS)	50 (6%)
Past Drug Dependence	91 (11%)	Schizoaffective	23 (3%)
Dysthymia	86 (11%)		
Post-Traumatic Stress Disorder	69 (9%)	Personality Disorder (Axis II)	149 (19%)

3. IRT Calibration Analysis

As previously noted, the MASS consists of 4 subscales, and a total of 616 items were administered out of the 626 available items. Overall internal consistency reliability (KR-21) was .814, which is excellent. KR-21 reliabilities for the separate scales were: MOODS, .49; OBS, .52; PAS, .40; SHY, .64. These findings support the idea that the items are tapping a primary core dimension. Correlations among the scales and the core affective domain ranged from .4 to .9. We next fit both unidimensional and bi-factor models (using the subscales as secondary dimensions) to the 616 item responses obtained from the 800 subjects. Both models converged

easily, which is reassuring given that no single subject took more than 154 items, yielding a sparse data matrix with large amounts of missing data. Nevertheless, the BIB design yielded a sufficient amount of overlap between the 616 items, so that both traditional unidimensional IRT and Bi-factor models could be fitted to the entire set of 616 items without computational difficulty. The Bi-factor model provided a significant improvement in fit to the observed item responses over the traditional unidimensional IRT model (chi-square = 2955; df=616, $p < .0001$).

Figure 1 presents several statistical features of the analysis results. Panel 1 presents a histogram of the estimated item thresholds, which describe the likelihood of a positive response conditional on level of impairment. Items with large *positive* thresholds are only rated positively by the most severely impaired subjects, whereas items with large *negative* thresholds are rated positively by subjects with both low and high levels of impairment. Panel A reveals that the scale provides excellent coverage of the impairment distribution. Panel B displays a histogram of the discrimination parameters for the primary dimension (expressed as factor loadings). Using a criterion of good discrimination equal to a factor loading of .3 or greater, approximately 90% of the items are good discriminators of high and low impairment, and are suitable for establishing the item bank.

Panel C displays a scattergram of the relationship between observed and expected (based on the bi-factor model) proportion rated positively for all 616 items. In general the bi-factor model provides an excellent fit to the observed data. Panel D displays the information function for the test as a whole. The curve is reasonably symmetric and Gaussian in appearance, again indicating that the test provides good coverage at both ends of the impairment spectrum. Panel E presents estimated reliabilities for the test as a whole (based on a block of 154 items taken by any single subject), where reliability is computed as $1 - 1/\text{information}$. Reliability is 0.95 or greater across the entire range of the impairment spectrum. Finally, Panel F presents four item characteristic curves, selected to illustrate items with high and low discrimination and high and low thresholds. The four items were:

Moods 82: *In the course of your life, including when you were a child, have you ever had periods of at least 3-5 days in which you were preoccupied with yourself and your own*

problems, thoughts and feelings?

OBS 23: *Did problems like being an obstinate or stubborn child, spending most of your time with your collections, needing to find just the right word or the exact pronunciation, or having tics or stuttering interfere with things you did outside of school?*

PAS 103: *In order to cope with the symptoms listed above (all panic symptoms listed in a box), did you need to take a walking stick or umbrella with you?*

SHY 135: *When attending or giving a party or meeting your friends, have you often felt embarrassed or uncomfortable?*

MOODS item 82 had high discrimination (factor loading = .70), and a low threshold (-.75) indicating that it has good discrimination of high and low levels of impairment, but is rated positively in people having relatively low levels of impairment. This result makes sense in that self-preoccupation is related to mood disorders, but is relatively common among people with high and low impairment. OBS item 23 also had high discrimination (factor loading = .63) but had a high threshold (.78) indicating that it is only rated positively by the most severely impaired subjects, as one would expect for an item related to stuttering and/or tics, which are less prevalent in the population. PAS item 103 has poor discrimination (factor loading = .20) and high threshold (1.28). This seems like an item indicative of quite unusual and severe pathology and it is therefore not surprising that it is not related to the core impairment dimension and is rarely rated positively. PAS 103 is an example of an item that would be eliminated from the item bank. SHY item 135 also has poor discrimination (factor loading = .14), but a low threshold (-.59). Being embarrassed or uncomfortable in social situations is quite common in general, and appears not to be related to the core mood disorder dimension. Again, this item would not be included in the item bank because it does not discriminate high and low levels of impairment.

These four items nicely illustrate the use of the bi-factor model in constructing an item bank for CAT administration. The purpose of the initial calibration is to determine which items are good

discriminators of the underlying dimension of interest. In our example, two items are good discriminators and two items are not. The items with poor discrimination would be either discarded, or sent back to the item developer to insure that the wording of the item correctly portrayed the intent of the question. The revised item might then be retested and if its discrimination was satisfactory, it might then be added to the item bank.

The results described here are estimates of the primary dimension of interest (i.e., mood disorder) adjusting for the conditional dependencies observed from the clustering of items within the four domains. This result is the distinct advantage of the bi-factor model and our example clearly illustrates its utility. Despite the multidimensional nature of the MASS we will be able to construct a single CAT and estimate a single mood disorder scale, without violating the assumptions of the statistical model, using the bi-factor model.

4. Comparison of Complete Responses versus Simulated CAT

To compare CAT versus complete test administration paradigms, we administered the complete 624 item MASS in two samples; 148 depressed patients from an earlier study conducted jointly by the Universities of Pittsburgh and Pisa (Study A), and 204 depressed patients collected as a part of the MHCAT study at the University of Pittsburgh (Study B). The data obtained from this administration were then used to compare full test administration (all 624 items) to a simulated CAT. The simulated CAT was obtained using the POSTSIM module of FastTEST PRO. Using this module we are able to take complete test data and simulate what would have happened if only a subset of responses had been obtained using CAT. Based on the standard CAT test administration, an initial item was selected and based on the actual response to that item, the next item was selected, and so on. Using a convergence criterion of a standard error of the mean (SEM) of 0.3, we are then able to determine on average (and range) how many items were required for CAT administration in each of the two samples that took the complete test. In addition, we compared simulated CAT to the complete test results in terms of bias (average difference between CAT and complete impairment estimates), and precision (average absolute difference).

Results of the analyses revealed that on average, 25 items in Study A (range of 18-55 items), and 24 items in Study B (range of 18-77 items) were needed to achieve convergence. This represents a savings of 96% over the complete test administration. The overall correlation between the CAT administered items and the total test score was $r=0.92$ for Study A, and $r=0.93$ for Study B. In terms of bias, the average difference was -0.277 for Study A and -0.095 for Study B. These results reveal that on average, there is negligible bias, with the CAT impairment estimates being slightly lower than the total impairment estimates. In terms of precision, the average absolute difference between CAT and complete test administration was 0.375 for Study A, and 0.371 for Study B. These results are consistent with the 0.3 SEM convergence criterion used in terminating the CAT, and represent a high level of precision. Figure 2 presents an illustration of an item by item CAT administration of the MASS resulting in a $SEM < .3$ in 26 items.

DISCUSSION

Mental health CAT (MHCAT) provides a new paradigm for conceptualizing and operationalizing psychiatric measurement. Rather than selecting a small number of “optimal” items, MHCAT begins with the establishment of a large item bank, consisting of as many as a thousand items. Based on a large calibration sample of respondents, the statistical characteristics of the items are investigated and the final item bank is reduced to the subset of items that are capable of good discrimination of high and low levels of the psychiatric dimension of interest. Deviations from unidimensionality of the resulting item bank can either be accommodated by the construction of several scales, or when items are sampled from domains of a common underlying trait, using a bi-factor model, which allows for within-domain residual association among item pairs. From this large item bank, a small subset of items are then administered adaptively, based on responses to the administration of previous items, until the uncertainty in the estimated impairment level is sufficiently small. In this way, mental health measurement scale developers can provide the broadest characterization of the construct in question (e.g., depression or mood disorder) but only administer those items that are relevant to the impairment level of the particular patient in question. This paradigm represents a distinct shift from the standard approach of administering small fixed-length tests, whose information content may be quite limited for subjects with differing levels of impairment.

Application of this statistical/psychometric methodology to a mental health testing problem yielded excellent results. First, the bi-factor model was shown to provide a quite reasonable representation of what was considered to be a test of high dimensionality that had previously been scored on four or more sub-domains of interest. This finding allowed us to make full use of all items in constructing the item bank which can be tapped to provide adaptive administration for an individual subject. Second, simulated CAT administration for subjects that had completed the entire test, revealed huge savings in the number of items administered for CAT relative to traditional scale administration. Indeed, a 96% reduction in items needed for accurate characterization of the mood disorder impairment level was observed for the CAT simulated from the full test administration data. The correlation between total MASS score and CAT-MASS score was $r=.92$ and $r=.93$ in two independent samples, indicating excellent agreement between the full test and reduced test administrations of the MASS. Bias was negligible, and precision quite high and consistent with the convergence criterion ($SEM=0.3$).

The primary limitation of this study is that it is based on the results obtained for a single rating scale of mood disorders. Whether the fit of the bi-factor model will be as good in other areas of psychiatric measurement remains an open question. Similarly, the large reduction in numbers of items needed for successful CAT administration (from 616 to 24) may also differ based on the specific domains of psychiatric illness under study. To this end, we are currently studying the application of IRT-based CAT to the problem of measuring major depression.

Figure 1 Panel A: Histogram of Estimated Item Thresholds
Positive Means that the Symptom is Present only in the Most Impaired Subjects

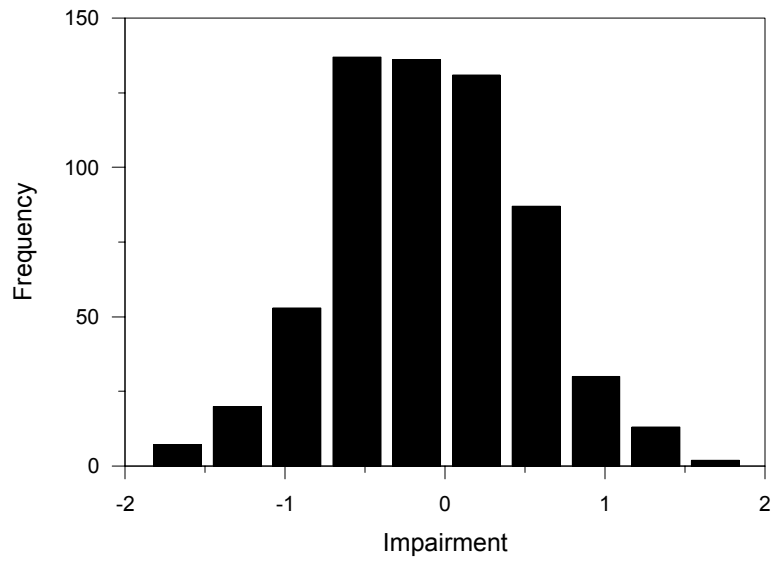


Figure 1 Panel B: Histogram of Primary Item Factor Loadings
Values of 0.3 or More Indicate Symptoms with Good Discrimination

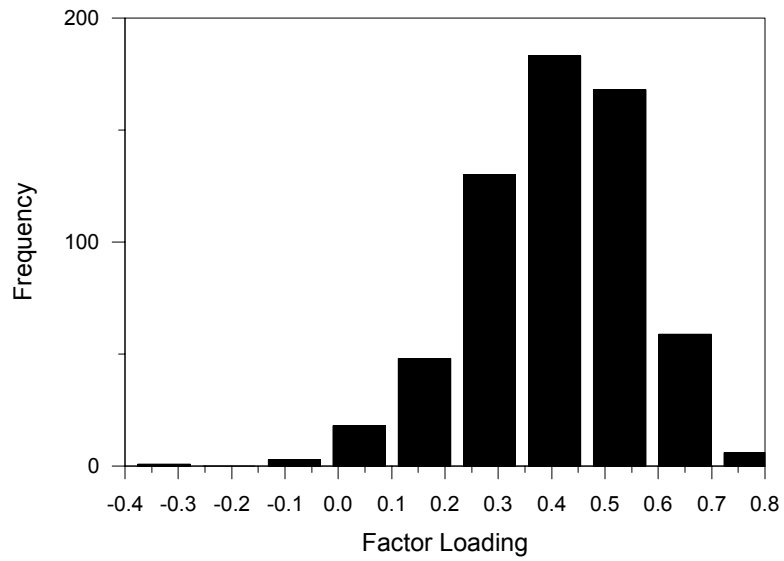


Figure 1 Panel C: Observed Versus Estimated Proportions
All 616 Items



Figure 1 Panel D: Test Information Function
Average Information Curve for a Block of 154 Items

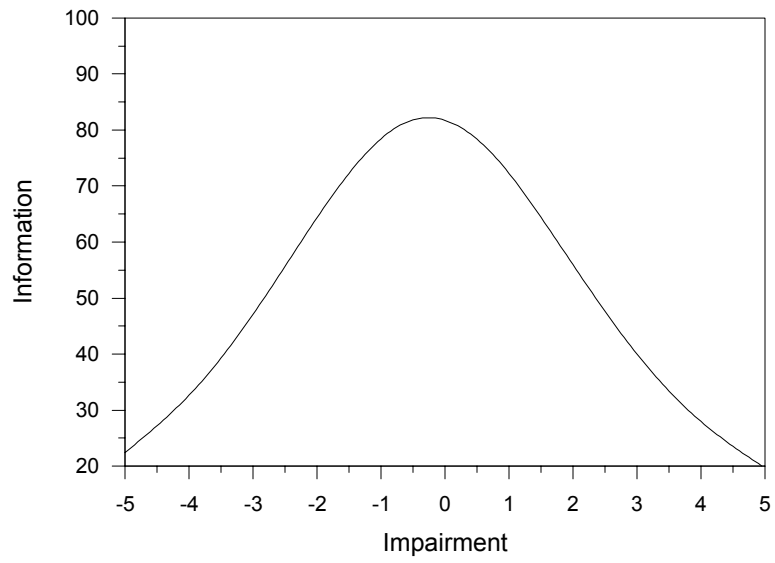


Figure 1 Panel E: Reliability Function for Average Block of 154 Items
Reliability - 1-1/Information

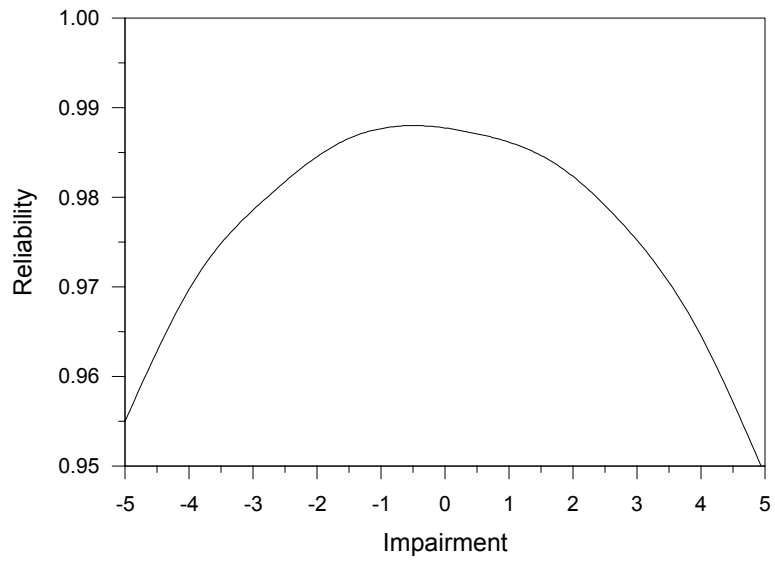


Figure 1 Panel F: Item Characteristic Curves for 4 Example Items
Items with high and low discrimination (d) and high and low thresholds (t)

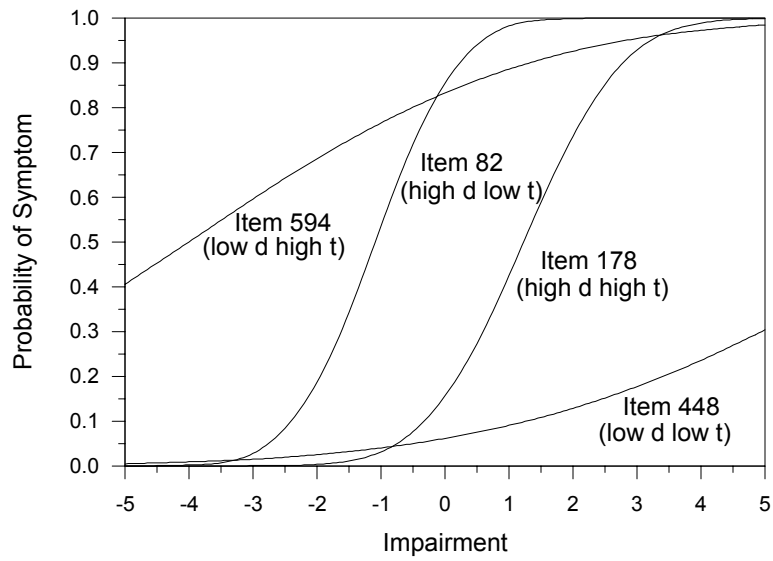


Figure 2: Item-By-Item Report of Maximum Information CAT

Theta was estimated by Bayesian EAP with a prior mean of 0.00
 and a prior variance of 1.00.
 The standard error band plotted is plus or minus 2.0 standard errors.
 X= Initial theta value Y = Yes N = No (or other response)

Item	Theta	SE	-3.....-2.....-1.....0.....+1.....+2.....+3
0	0.00	1.000	-----X-----
1	0.52	0.813	-----Y-----
2	0.95	0.708	-----Y-----
3	1.24	0.654	-----Y-----
4	1.42	0.622	.-----Y-----
5	1.19	0.544	.-----N-----
6	1.33	0.524	.-----Y-----
7	1.17	0.477	.-----N-----
8	1.26	0.462	.-----Y-----
9	1.38	0.454	.-----Y-----
10	1.29	0.428	.-----N-----
11	1.09	0.398	.-----N-----
12	1.15	0.388	.-----Y-----
13	1.22	0.382	.-----Y-----
14	1.11	0.363	.-----N-----
15	1.17	0.358	.-----Y-----
16	1.23	0.354	.-----Y-----
17	1.28	0.350	.-----Y-----
18	1.36	0.349	.-----Y-----
19	1.42	0.347	.-----Y-----
20	1.49	0.346	.-----Y-----
21	1.54	0.344	.-----Y-----
22	1.49	0.333	.-----N-----
23	1.40	0.321	.-----N-----
24	1.34	0.311	.-----N-----
25	1.37	0.308	.-----Y-----
26	1.26	0.297	.-----N-----
Item	Theta	SE	-3.....-2.....-1.....0.....+1.....+2.....+3

The final theta estimate based on 26 items was 1.26 with a standard error of 0.297, resulting in a 2.00 standard error band of 0.67 to 1.86

REFERENCES

1. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*. 1981; 46: 443-459.
2. Bock RD, Gibbons RD, Muraki E. Full-information item factor analysis. *Applied Psychological Measurement*. 1988;12: 261-280.
3. Gibbons RD, Hedeker D. Full-information item bi-factor analysis. *Psychometrika*. 1992;57: 423-436.
4. Weiss DJ. Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*. 1985; 53: 774-789.
5. Kingsbury GG, Weiss DJ. *An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability tests*. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory; 1980: Research Report 80-5.
6. Kingsbury GG, Weiss DJ. A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In: Weiss DJ, ed. *New horizons in testing: Latent trait theory and computerized adaptive testing*. New York: Academic Press; 1983: 257-283.
7. McBride JR, Martin JR. Reliability and validity of adaptive ability tests in a military setting. In: Weiss DJ, ed. *New horizons in testing: Latent trait theory and computerized adaptive testing*. New York: Academic Press; 1983: 223-236.
8. Baker FB. *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker, Inc; 1992.
9. Weiss DJ, McBride JR. Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement*. 1984; 8: 272-285.
10. Hambleton RK, Swaminathan H. *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff; 1985.
11. Weiss DJ, Kingsbury GG. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*. 1984; 21: 361-375.
12. Vale CD, Weiss DJ. *A rapid item-search procedure for Bayesian adaptive testing*. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory; 1984: Research Report 77-4.
13. Assessment Systems Corporation. *Manual for the MicroCAT Testing System*: Third Edition. St. Paul, MN; 1987.

14. Assessment Systems Corporation. *The FastTEST Professional Testing System*. St. Paul, MN; 2000.
15. Brown JM, Weiss DJ. An adaptive testing strategy for achievement test batteries. Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing Laboratory; 1977: Research Report 77-6.
16. Andrich D. A rating formulation for ordered response categories. *Psychometrika*. 1978a: 43: 561-571.
17. Andrich D. Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*. 1978b: 2: 581-594.
18. Andrich D. The application of an unfolding model of the IRT type to the measurement of attitude. *Applied Psychological Measurement*. 1988: 12: 33-51
19. Muraki E. Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*. 1990: 14: 59-71.
20. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969: 17: 1-68.
21. Reise SP, Waller NG. Fitting the two-parameter model to personality data. *Applied Psychological Measurement*. 1991: 15: 45-58.
22. Baek SG. Computerized adaptive testing using the partial credit model for attitude measurement. In: Wilson M, Engelhard Jr. G, & Draney K., eds. *Objective measurement: Theory into practice*. Volume 4. Norwood NJ: Ablex; 1997.
23. Dodd BG, DeAyala RJ, Koch WR. Computerized adaptive testing with polytomous items. *Applied psychological measurement*. 1995: 19: 5-22.
24. Thurstone LL. *Multiple Factor Analysis*. Chicago, University of Chicago Press; 1947.
25. Holzinger KJ, Swineford F. The bi-factor method. *Psychometrika*. 1937: 2: 41-54.
26. Tucker LR. An inter-battery method of factor analysis. *Psychometrika*. 1958: 23: 111-136.
27. Joreskog KG. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*. 1969: 34: 183-202.
28. Muth'en BO. Latent variable modeling in heterogeneous populations. *Psychometrika*. 1989: 54: 557-585.
29. Gibbons RD, Bock, RD, Hedeker D, Weiss D, Bhaumik D, Kupfer D, Frank E, Grochocinski V, Stover A. Full-information item bi-factor analysis of graded response data. 2005, submitted.

30. Likert R. A technique for the measurement of attitudes. *Archives of Psychology*. 1932: Monograph 140.
31. Cassano GB, Michelini S, Shear MK, Coli E, Maser JD, Frank E. The panic-agoraphobic spectrum: A descriptive approach to the assessment and treatment of subtle symptoms. *American Journal of Psychiatry*. 1997; 154(suppl 6): 27-38.
32. Frank E, Cassano GB, Shear MK, Rotondo A, Dell'Osso L, Mauri M, Maser J, Grochocinski VJ. The spectrum model: A more coherent approach to the complexity of psychiatric symptomatology. *CNS Spectrums*. 1998; 3, 23-34.
33. Cassano GB, Banti S, Mauri M, Dell'Osso L, Miniati M, Maser JD, Shear MK, Frank E, Grochocinski VJ, Endicott J. Internal consistency and discriminant validity of the Structured Clinical Interview for Panic-Agoraphobic Spectrum (SCI-PAS). *International Journal of Methods in Psychiatric Research*. 1999a; 8: 138-145.
34. Cyranowski JM, Shear MK, Rucci P, Fagiolini A, Frank E, Grochocinski VJ, Kupfer DJ, Banti S, Armani A, Cassano GB. Adult separation anxiety: Psychometric properties of a new structured clinical interview. *J Psychiatr Res*. 2002; 36: 77-86.
35. Dell'Osso L, Cassano GB, Sarno N, Millanfranchi A, Pfanner C, Gemignani A, Maser JD, Shear MK, Grochocinski VJ, Rucci P, Frank E. Validity and reliability of the Structured Clinical Interview for Obsessive-Compulsive Spectrum (SCI-OBS) and of the Structured Clinical Interview for Social Phobia Spectrum (SCI-SHY). *International Journal of Methods in Psychiatric Research*. 2000; 9: 11-24.
36. Fagiolini A, Dell'Osso L, Pini S, Armani A, Bouanani S, Rucci P, Cassano GB, Maser JD, Endicott J, Shear MK, Grochocinski VJ, Frank E. Validity and reliability of a new instrument for assessing mood symptomatology: The structured clinical interview for mood spectrum (SCI-MOODS). *The International Journal of Methods in Psychiatric Research*. 1999; 8: 71-82.
37. Mauri M, Borri C, Baldassari S, Benvenuti A, Rucci P, Cassano GB, Shear MK, Grochocinski VJ, Maser JD, Frank E. Acceptability and psychometric properties of the Structured Clinical Interview for Anorexic - Bulimic Spectrum (SCI-ABS). *Int J Methods Psychiatr Res*. 2000; 9: 68-78.
38. Sbrana A, Dell'Osso L, Gonnelli C, Impagnatiello P, Doria MR, Spagnolli S, Ravani L, Cassano GB, Frank E, Shear MK, Grochocinski VJ, Rucci P, Maser JD, Endicott J. Acceptability, validity and reliability of the Structured Clinical Interview for the Spectrum of Substance Use (SCI-SUBS): A pilot study. *Int J Methods Psychiatr Res* 2003; 12:105-115.
39. Sbrana A, Dell'Osso L, Benvenuti A, Rucci P, Cassano P, Banti S, Gonnelli C, Doria MR, Ravani L, Spagnolli S, Rossi L, Raimondi F, Catena M, Endicott J, Frank E, Kupfer DJ, Cassano GB. The psychotic spectrum: Validity and reliability of the Structured

Clinical Interview for the Psychotic Spectrum. *Schizophr Res*. Submitted.

40. Dell'Osso L, Armani A, Rucci P, Frank E, Fagiolini A, Corretti G, Shear MK, Grochocinski VJ, Maser JD, Endicott J, Cassano GB. Measuring mood spectrum disorder: Comparison of interview (SCI-MOODS) and self-report (MOODS-SR) instruments. *Compr Psychiatry*. 2002a: 42:69-73.
41. Shear MK, Frank E, Rucci P, Fagiolini A, Grochocinski VJ, Houck P, Cassano GB, Kupfer DJ, Endicott J, Maser JD, Mauri M, Banti S. Panic-agoraphobic spectrum: Reliability and validity of assessment instruments. *Journal of Psychiatry Research*. 2001: 35: 59-66.
42. Dell'Osso L, Rucci P, Cassano GB, Maser JD, Endicott J, Shear MK, Sarno N, Saettoni M, Grochocinski VJ, Frank E. Measuring social phobia and obsessive-compulsive disorders: Comparison of interviews and self-report instruments. *Comprehensive Psychiatry*. 2002b: 43: 81-87.
43. Rucci, P. [No relationship between total PAS-SR score and Hamilton Rating Scale for Depression (HRSD) score at the time of PAS-SR administration.]. Unpublished raw data; 2004.
44. Frank E, Shear MK, Rucci P, Banti S, Mauri M, Maser JD, Kupfer DJ, Miniati M, Fagiolini A, Cassano GB. Cross-cultural validity of the Structured Clinical Interview for Panic-Agoraphobic Spectrum. *Soc Psychiatry Psychiatr Epidemiol*. Submitted.
45. Rucci P, Cassano GB, Frank E, Fagiolini A, Dell'Osso L, Shear MK, Kupfer DJ. The mood spectrum in unipolar and bipolar patients. *Bipolar Disord*. 2003a: 1: 77-78.
46. Cassano GB, Dell'Osso L, Frank E, Miniati M, Fagiolini A, Shear K, Pini S, Maser J. The bipolar spectrum: A clinical reality in search of diagnostic criteria and an assessment measure. *J Affect Disord*. 1999b: 54: 319-328.
47. Cassano GB, Rucci P, Frank E, Fagiolini A, Dell'Osso L, Shear MK, Kupfer DJ. The mood spectrum in unipolar and bipolar disorder: Arguments for a unitary approach. *Am J Psychiatry*. 2004: 161:1264-1269.
48. Mundt JC, Marks IM, Shear MK, Greist JH. The Work and Social Adjustment Scale: A simple measure of impairment in functioning. *Br J Psychiatry*. 2002: 180:461-464.
49. Cochran WG and Cox GM. *Experimental Designs*. New York, Wiley; 1957.